图书情报工作
第 64 卷 第 8 期 2020 年 4 月

ChinaXiv合作期刊

Prophet 预测 - 修正的主题强度演化模型
——以干细胞领域为实证*

■ 张鑫 文奕 许海云 刘忠禹
中国科学院成都文献情报中心 成都 610041

摘 要:[目的/意义]主题演化对科技前沿探测、创新战略部署具有十分重要的作用。[方法/过程]将主题演化分析过程分解为主题的表示、相似性关联和强度演化计算几个步骤,提出一种主题强度演化与预测模型,使用 LDA 模型进行主题的表示,提出内容、共现和趋势相似度等维度进行主题关联计算,引入基于 Prophet 的预测 - 修正模型进行主题演化趋势预测。并以干细胞领域为例,进行演化的实证分析。[结果/结论]实验表明,对每个研究主题采用 Logistic 增长模型进行预测 R2Score 都达到 0.90 以上,表明 Prophet 中的 Logistic 增长模型与该领域主题增长趋势规律相符合,能够较好地拟合主题强度的演化趋势。提出的主题演化模型对专业领域内主题分布与演化分析有一定的借鉴意义。

关键词: 主题演化 主题相似性 时间序列 Prophet
分类号: G251
DOI: 10.13266/j.issn.0252 - 3116.2020.08.010

chinaXiv:202304.00256v1

研究主题演化是对研究主题的产生、扩散和发展过程的挖掘、分析和展示。它能够帮助情报分析和科技管理人员全面、客观地把握领域的创新发展规律,因而成为科技前沿探测、技术预见以及科技路线图制定中的基础性、核心性工作。深刻认识并准确把握领域科技创新规律与演化趋势,系统谋划创新发展新路径,对于科技前沿预测、创新战略部署具有至关重要的作用。在大数据时代,科技文献数量爆发式增长,通过大规模知识计算理论方法对海量科技文献进行深度、自动化的加工和挖掘,成为目前研究主题演化的主流方法。

1 研究现状

研究主题演化和创新前沿预测是情报学关注的基本问题之一,在情报学初创时期已经被提出来,经历了 40 余年的发展,不断有新的思想、新方法融入其中,目前主要的研究方法有基于专家知识的方法、基于引文的方法和基于文本挖掘的方法等几类。也有学者将研究主题演化分为基于定性研究、基于定量研究、定性定量相结合 3 类方法。基于专家知识的方法主要属于定性研究范畴,基于引文和文本挖掘的方法主要属于定量研究范畴。

基于专家知识的方法。传统的学科研究主题识别主要依靠专家知识进行判读,主要使用专家访谈方法、德尔菲法、TRIZ 方法、形态分析方法[1]等,这些方法主观性较强,而且成本较高,但由于领域专家的公信力较好,这些方法也是目前被广泛采用的,准确率最好的研究方法。

基于引文的方法。由于引用信息能够很好地表示知识的传承信息,引用信息在研究主题发现和演化分析上有非常重要的作用,此类方法主要有 N. P. Hummon[2]、A. Martinelli[3]、L. Y. Y. Liu 等[4]等使用的引文主路径方法和 A. Pilkington 等[5]、R. J. Lai 等[6]使用的引文聚类方法等。

基于文本挖掘的方法。随着深度学习和自然语言处理等技术的发展和计算机处理能力的提升,文本挖

* 本文系国家自然科学基金项目“基于科学 - 技术主题关联分析的创新演化路径识别方法研究”(项目编号:71704170)和中国科学院信息化专项“面向干细胞领域知识发现的科研信息化应用”(项目编号:XXH13506 - 203)研究成果之一。
作者简介:张鑫(ORCID:0000 - 0001 - 8784 - 3788),馆员,硕士;文奕(ORCID:0000 - 0002 - 6520 - 2733),研究馆员,硕士,通讯作者,E-mail:wenyi@clas.ac.cn;许海云(ORCID: 0000 - 0002 - 7453 - 3331),副研究员,博士; 刘忠禹(ORCID:0000 - 0003 - 3852 - 1947),助理馆员,硕士。
收稿日期:2019 - 07 - 28 修回日期:2019 - 11 - 07 本文起止页码:78 - 92 本文责任编辑:徐健

ChinaXiv合作期刊

掘方法在研究主题演化分析中,发挥着越来越重要的作用。文本挖掘的方法又可以分为:基于关键词共现的方法[7]、基于句法结构分析的方法[7]和基于概率主题模型的方法[8-10]等。

3 种主题演化分析方法的对比如表 1 所示:

表 1　研究主题演化分析的主要方法

方法	主要思路	优势	当前不足
基于专家知识的方法	通过专家访谈等方法建构领域研究主题演化趋势	专家解读权威性	依靠人工,花费较大
基于引文分析	文献的引用关系计算文献及主题相似性	多种引文关系定量描述文献之间的相似性,有利于发现主题的继承发展关系	引文覆盖面较窄,引文动机干扰,有时滞性
基于文本挖掘	通过计算主题的词分布,及分布距离得到主题相似性	极大地促进了主题演化分析方法的自动化与效率,有助于主题相关性测度	依赖计算机处理,运算环节对结果影响较大

目前的主题演化分析方法重现状分析、轻未来预测。特别是,最近时间片内的主题趋势分析不准确,这主要是由于论文发表时延或数据出版商收录数据的时延等因素,使得最近时间片内论文收录不完全导致的。笔者认为,最近时间片内的论文数据有如下两个特点:

(1)数据价值高。最近时间片内论文与现在时间最近,最能反映近期的研究主题的分布情况,数据价值比较高,不能完全舍弃;

(2)数据不完整。最近的论文数据又是不完整的,如果只是使用这个不完整的数据,进行展示和预测,会产生不正确的分析和预测结果。

2　研究方法流程

本文的研究主题演化的研究流程整体分为两个阶段:第一阶段为数据处理与研究主题识别阶段,其核心是研究主题表示与抽取;第二阶段为研究主题趋势分析与展示阶段,其核心是主题关联与主题趋势预测。两者的关系为主题表示与抽取是后续趋势分析的基础和前提,一个好的主题表示方法才能使得后面趋势分析结果站得住脚、容易解释,主题趋势分析是目的和结果,趋势分析的结果能够服务于科研态势分析、科技决策部署等情报分析任务。具体如图 1 所示:

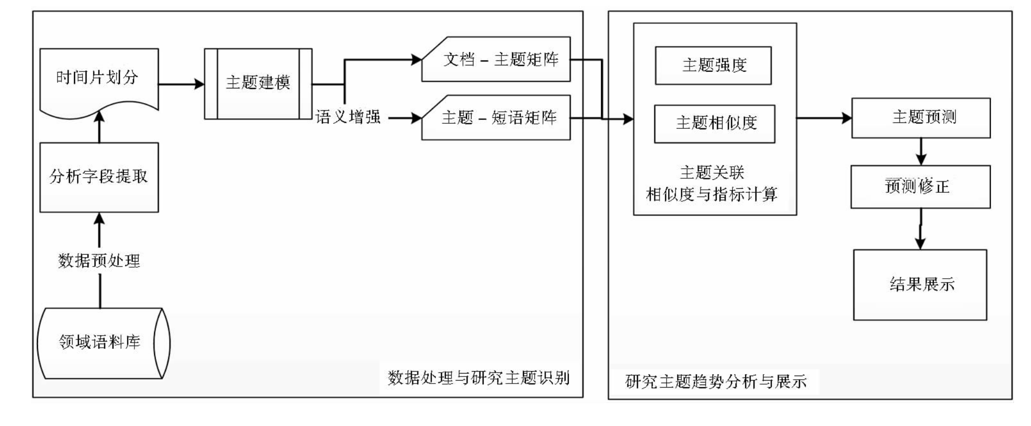


图 1　主题演化分析的方法流程

(1)数据处理与研究主题识别。从数据库提供商处,采用一定的检索策略进行数据检索,将下载数据进行数据去重、缺失项处理、停用词去除和词干还原等操作,形成清洗过后的领域语料库。抽取待分析字段(关键词、标题、摘要或全文等),再按照一定的规则进行时间片划分。用 2.1 节介绍的主题建模方法进行主题抽取与主题语义增强,得到文档主题关系和主题－词语关系两个矩阵。

(2)研究主题趋势分析与可视化展示。对计算得到的研究主题,采用不同的相似度计算方式进行主题关联,在不同时间片内进行主题强度计算,得到相应的时间序列,再在数据完整准确的阶段采用 2.3 节中叙述的时间序列分析方法进行主题趋势预测,使用近期不完整的数据进行预测修正,获得主题趋势。以研究主题的生命周期理论为指导,并与领域专家向结合,对主题趋势进行分析和解读。以折线图、主题河流图等形式可视化展示出研究主题强度变化趋势。

2.1　主题表示建模

LDA 主题模型是 D. M. Blei 等在 2003 年提出的主题表示模型[9],由于它能够很好地抽取出文档中的隐性主题,迅速成为主题抽取与表示领域使用最为广

chinaXiv:202304.00266v1

ChinaXiv合作期刊

泛的模型,后面很多人对原始的LDA模型进行了各种各样的改进,如DTM[11]、TOT[12]等,但后面一些算法的计算复杂度更高,不容易在情报分析工具中进行集成。

LDA模型是一种贝叶斯概率模型,它假设文档是由若干的隐含主题构成的,而主题是由词构成的。具体而言,假设有文档集D,其中有M篇文档$d_1,d_2,\cdots,d_M$,第m篇文档的长度为N_m,则LDA模型的文档生成过程为:①从参数为$\vec{\alpha}$的Dirichlet分布中采样生成文档d_i的主题分布θ_i;②从参数为θ_i的多项式分布采样生成文档d_i中第j个单词$w_{i,j}$的主题$z_{i,j}$;③从参数为$\vec{\beta}$的Dirichlet分布中采样生成主题$z_{i,j}$的词语分布$\Phi_{z_{i,j}}$;④从词语的多项式分布$\Phi_{z_{i,j}}$中采样最终生成词语$w_{i,j}$,设文档集中词典的大小为N_d,主题的个数为N_T(下面几节符号相同)。

这样,单词和主题联合概率分布可以表示为

$$p(\vec{w},\vec{z}\mid\vec{\alpha},\vec{\beta})=\prod_{k=1}^{K}\frac{\Delta(\vec{n_k}+\vec{\beta})}{\Delta(\vec{\beta})}\prod_{m=1}^{M}\frac{\Delta(\vec{n_m}+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

公式(1)

可以利用公式(1)进行主题模型参数估计,D. M. Blei原始论文中采用E-M方法进行参数估计,运行效果较慢。I. Porteous等[13]提出了Collapse Gibbs-Sample方法,大大加快了主题模型的训练速度,加速了主题模型的落地应用。M. Hoffman等[14]提出了LDA模型的Online Learning方法,采用批量更新,再合并的方式进行训练,是大数据下主题训练成为可能。广泛使用的基于Python的Gensim工具箱中使用的就是这种方法。如图2所示:

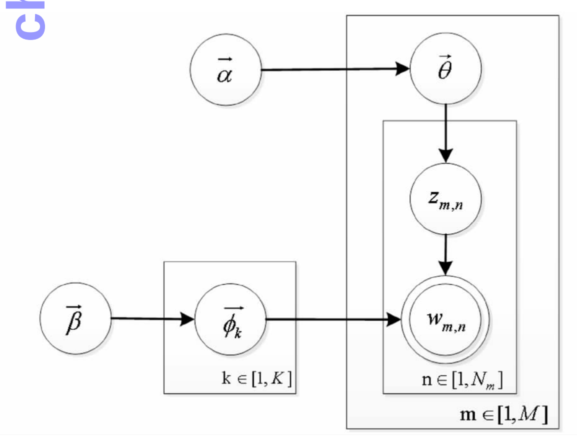


图2 LDA主题模型

主题模型在应用过程中,为了更好的表示实际模型,有两个问题有待讨论:

(1)主题模型参数选择问题。主题模型在训练的过程中需要指定的参数主要有α,β和主题个数K等,特别是关于最优主题个数的选择问题,可以使用困惑度(perplexity)指标[9]指导主题个数选择,后面的HDP[15]等非参数主题模型,虽然引入了层次特征,自动进行选择,但计算复杂度比较高,效果也不是很理想。最近的研究一般以Topic Coherence等指数进行主题评价,如D. Mimno等[16]。还有一些研究使用融合特征进行主题个数选择,如王婷婷等[17]。笔者采用困惑度和Coherence两个指标相结合进行主题个数选择。这种方法相对来说执行速度快,抽取效果较理想,且需要的附加特征比较少。这种方法给出的只是给出主题个数建议,最终主题个数还要结合抽取的结果,由专家判读决定。

(2)主题的语义增强问题。传统的主题模型采用一组单词进行主题表示,往往不容易进行主题解读,针对此问题,很多学者提出了一系列的语义增强方法,如TNG[18]、CITPM[19]、PhraseLDA[20]、Chunk-LDAvis[21]等模型,这些模型都使用词组进行主题表示,可读性更强。笔者采用Bi-Gram进行主题语义增强,使得抽取的主题中不仅含有单词,还有常见Bi-Gram短语信息,这种方法实现速度快,人工干预少,且只需要在数据处理时进行一次加工,不需要在抽取结果后进行二次加工替换操作,在大规模预料抽取上性能效率高于其他方法。

2.2 主题关联

2.2.1 主题强度

在主题抽取之后,需要对主题强度进行计算。主题强度是主题本身具有的统计属性,用来表征主题受关注的程度。当前的话题计算方法主要采用基于主题支持的文档数量、基于语料库主题概率、基于文本的显著性3种,孙孟孟[22]等比较了3种方法,得到结论3种方法在长文本下能够得到一致性的分析结果,而方法一得到的结果相对更加显著,故本文采用方法一进行主题强度表征。

主题强度定义:在第u个时间片内,文档集合中文档数量为D_u,主题j的强度可以定义为归属为主题j的文章数目

$$ST_j=\sum_{d\in D_u}\theta_{dj}$$

公式(2)

2.2.2 主题相似性度量方法

主题抽取之后,笔者希望探求各个主题之间的关联关系。本文中的关联关系采用主题相似性进行描述,传统的主题相似性从主题内容分布维度进行计算。除此之外,笔者采用主题共现、时间趋势两个新的视角进行相似性度量,提出共现相似性、趋势相似性指标,并对这3种相似性度量方式进行一致性分析。

chinaXiv:202004.00266v1

（1）主题内容相似性。主题的内容相似性，用来表征各个研究主题在内容结构上的相似性。具体而言，是用主题在词语表示分布上的相似性来度量主题相似性，表示分布相似性的方式有很多，如 Kullback-Leibler (KL)散度、Hellinger 距离、Jaccard 距离、Jensen-Shannon (JS)散度等。JS 散度由于具有对称性的优点，更符合主题相似度计算场景，笔者选用 JS 散度作为内容相似性度量方式。设主题 $T_i, i \in [1, N_T]$，对词典中单词 $w_k, k \in [1, N_d]$ 的概率为 φ_{ik}，主题内容相似性计算公式如下：

$$simContent(T_i, T_j) = 1 - JS(\vec{\varphi_i}, \vec{\varphi_j}) = 1 - [\frac{1}{2} KL(\vec{\varphi_i} || \frac{\vec{\varphi_i} + \vec{\varphi_j}}{2}) + \frac{1}{2} KL(\vec{\varphi_j} || \frac{\vec{\varphi_i} + \vec{\varphi_j}}{2})] \quad \text{公式 (3)}$$

其中，KL 函数为 KL 散度，具体的计算公式为：

$$KL(\vec{\varphi_i} || \vec{\varphi_j}) = \sum_{k=1}^{N_T} \varphi_{ik} \log \frac{\varphi_{ik}}{\varphi_{jk}} \quad \text{公式 (4)}$$

（2）主题共现相似性。研究主题除了具有内容结构外，还有其他的一些属性特征，可以通过这些属性特征来刻画主题相似性，例如笔者提出通过研究主题在文档中共同出现的频次，来表征主题之间的共现相似性。设文档 $d_m, m \in [1, M]$ 中，主题 T_i 的概率为 θ_{mi}，主题 T_j 的概率为 θ_{mj}，则这两个主题在这篇文档中的共现度为 $\min(\theta_{mi}, \theta_{mj})$。两个研究主题在整个文档集合中的共现相似性为：

$$simCoocur(T_i, T_j) = \sum_{m=1}^{M} \min(\theta_{mi}, \theta_{mj}) \quad \text{公式 (5)}$$

（3）主题趋势相似性。主题的趋势相似性，用来度量不同的技术主题在时间演化趋势上的相似性。每个研究主题在不同时间片内的主题强度构成时间序列，我们自然希望通过这种时间序列之间的相似性来刻画主题在趋势维度上的相似性，基于此笔者提出趋势相似性定义为：

$$simTrends(T_i, T_j) = (1 + dist(T_i, T_j))^{-1} \quad \text{公式 (6)}$$

其中，$dist(T_i, T_j)$ 为主题 T_i 与主题 T_j 构成的时间序列 $\{ST_i\}_U$ 与 $\{ST_j\}_U$ 之间的距离度量。时间序列相似性度量方法有锁步类(lock-step)度量和弹性类(elastic)度量等，锁步度量主要有欧式距离、马氏距离等，弹性度量主要有动态时间规整(DTW)方法等，DTW 方法能够克服欧式距离方法的缺点，支持序列平移，灵活方便地处理多相位序列，是时间序列度量的最常用方法，DTW 方法的距离计算方法采用动态规划的方法，具体更新公式可以表示为：

$$dist(T_i, T_j) = comDist(T_{1i}, T_{1j}) + \min(dist(R(T_i), R(T_j)), dist(T_i, R(T_j)), dist(R(T_i), T_j)) \quad \text{公式 (7)}$$

式中 $R(T_i)$，表示 T_i 中剩余的序列，$comDist(T_{1i}, T_{1j})$ 表示两个序列中第一个时间点的距离，实际中可以选用欧式距离等度量。

2.2.3 三种主题相似性度量方法的一致性

三种相似性方法从不同的角度进行主题相似性度量，需要的数据和计算复杂性各异。那么同样是度量相似性，这三种方法得到的结果是不是一致的呢？笔者采用编辑距离的方法度量不同主题相似性度量方法的结果一致性。采用 2.2.2 节中介绍的相似的度量方法，按照和某个主题相似度从大到小的顺序，得到一个序列，通过比较不同主题相似性度量结果得到的序列的相似性进而衡量方法的一致性。笔者通过 Vladimir Levenshtein 提出的编辑距离来度量序列相似性，这种编辑距离表示从一个序列经过插入、删除或替换操作变换成另一个序列的最小操作数目。两种相似性度量方法生成序列的编辑距离就越小，说明它们的一致性越强，反之则说明它们一致性弱。

2.3 主题趋势预测

每个主题的强度形成一个时间序列，如果以一年为一个时间片，由于最近时间片数据不完整，首先去掉不完整数据时间片，用前面的数据进行建模。但同时最近时间片的数据更可以反应最近时间内的主题趋势，故使用这个时间片内的数据进行预测修正，并基于模型和修正数据进行未来的趋势预测。

问题描述：设第 u 个时间片内，主题 j 的强度为 ST_{uj}，对每个主题 j，不同时间片内的主题强度构成一个时间序列 $\{ST_{1j}, ST_{2j}, ST_{3j}, \cdots ST_{(U-1)j}\}$，希望预测第 T(T > U) 个时间片内的主题强度 $\widehat{ST}_{Tj}$。

（1）基本模型。目前主流的时间序列模型主要有 ARIMA 模型、LSTM 神经网络模型[23]等。ARIAM 模型在短时间预测比较有效，LSTM 在长时间的预测较为有效，具体到本问题中，由于以年为时间单位，训练数据较少，LSTM 类神经网络模型很难达到收敛。

2018 年 Facebook 开源了 Prophet(先知)神经网络预测工具[24]之后，其迅速成为时间序列分析的热门工具，截至 2019 年 7 月 22 日，项目在 Github 上的关注度：Watch(设置项目变化邮件提醒的人数)值为 376，Star(关注项目的人次)值为 8 888，Fork(拷贝项目的人次)值为 2 172。Prophet 模型是一个加法模型，它假设

chinaXiv:202004.00266v1

ChinaXiv合作期刊

观测变量的规律满足如下公式：

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad \text{公式 (8)}$$

其中，$g(t)$ 为非周期性的增长的趋势项，$s(t)$ 是周期因素项，$h(t)$ 为节假日因素项，ε_t 为满足正态分布的误差项。和以往的模型相比，Prophet 模型具有自动性好、可解释性强、可扩展性强、训练速度快等优点。对本研究而言，各个研究主题的发展呈现明显的增长趋势，而且数据点个数相对不多。相校于经典的 ARIMA 模型，Prophet 模型能够更好地预测增长趋势，且比 LSTM 等需要大样本数据训练的模型，更容易达到收敛。

Prophet 中使用了饱和增长（logistic）与分段线性（linear）两种趋势增长模型，此外 Prophet 模型还将变点检测显式引入模型中，若设模型随时间变化的承载能力为 $C(t)$，在 S 个变点处的变化率向量为 δ，计算处的调整向量为 γ，$a(t)$ 为指示向量 $\in\{0,1\}^S$，则有饱和模型趋势增长公式为：

$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^T\delta)(t - (m + a(t)^T\gamma))} \quad \text{公式 (9)}$$

线性模型的增长公式为：

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma) \quad \text{公式 (10)}$$

在本研究中，引入 Prophet 趋势增长模型进行主题趋势预测。具体做法为，以年为单位进行时间切片，暂时不考虑周期因素和节假日因素的影响，设置 weekly_seasonality = False，daily_seasonality = False。由于趋势数据量相对较少，将 changepoint 的个数设置小些，本研究中设置为 3，其他参数使用模型默认。

（2）预测修正。由于所有研究主题对应的文章的采集时间是相同的，在最后时间片内，可以近似地认为每个研究主题的缺失比例都是一样的。根据这个比例，笔者提出对 Prophet 模型预测出来的数值进行修正（见公式 11），ST_{Tj} 为实际观测到的 T 时刻主题强度，$\widehat{ST}_{Tj}$ 为采用 Prophet 模型预测出来的 T 时刻主题强度，ST^*_{Tj} 为修正后的 T 时刻主题强度。

$$ST^*_{Tj} = \frac{ST_{Tj} * \sum_{j=1}^{J}\widehat{ST}_{Tj}}{\sum_{j=1}^{J} ST_{Tj}} \quad \text{公式 (11)}$$

加入修正模型后，整体的预测流程变为 3 阶段模型：①去掉不完整数据，使用 Prophet 模型进行预预测；②根据最近时间片的数据，使用公式（8）对预预测结果进行数值修正；③基于预测 - 修正的数据进行后面时间片的主题强度演化趋势预测。

3 实证研究

干细胞与再生医学的研究为癌症等疾病的治疗带来革命性的变革，9 次入选美国《科学》杂志十大科技进展，也是当前国内外生物医学领域的研究热点，国家重点研发计划等重大科技项目中也多次布局相关项目，故笔者选取干细胞领域进行实证研究。在 ISI Web of Knowledge 中以检索式（TI = Stem Cells）进行检索，检索时间 2019 年 5 月，检索得到文章 433 469 篇，检索结果中不同年份的文章数量如图 3 所示，可以看出文章数量呈现出缓慢增长到快速增长再到饱和增长的趋势，最近两年由于数据收录不完整，故呈现下降趋势。

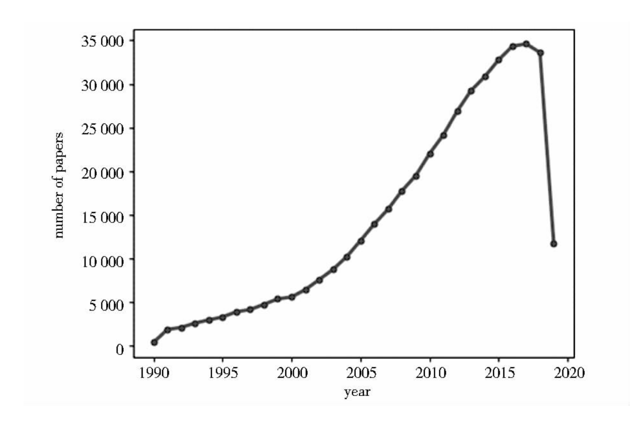


图 3　不同年份的文章数量

3.1 数据预处理

（1）关键词抽取。在数据预处理过程，首先提取出论文中的标题和摘要字段，其次使用正则表达式匹配的方式去掉原始文献中的标点符号、数字、email 地址等特殊字符，使用 gensim 工具中的 utils. simple_preprocess 工具进行初步的分词处理。之后，使用工业级的自然语言处理工具 Spacy 进行句子中单词的词性分析，抽取出抽取名词、动词、形容词、副词实词作为主题抽取的对象。

（2）主题的语义增强。主题的语义增强有在预处理时进行增强和主题抽取后进行增强两种方式，笔者采用文章[17]类似的方式，综合考虑计算时间复杂度，使用 Bi-Gram 进行增强，用 gensim 工具中的 models. Phrases 工具提取出 Bigram 短语，加入原始文本之中，使得提取出的主题中能够含有更多可解释性信息。

3.2 干细胞领域研究主题的抽取

采用 Gensim 中的 LDA 模型，参数 alpha 设置为 ‘auto’，采用困惑度和连贯性两个指标指导进行主题个数选择，困惑度指标计算采用包中的 log_perplexity

chinaXiv:202304.00206v1

ChinaXiv合作期刊

张鑫,文奕,许海云,等. Prophet 预测 - 修正的主题强度演化模型——以干细胞领域为实证[J]. 图书情报工作,2020,64(8):78 - 92.

进行计算,连贯性指标采用 Gensim 中的 models. coherencemodel 进行计算。

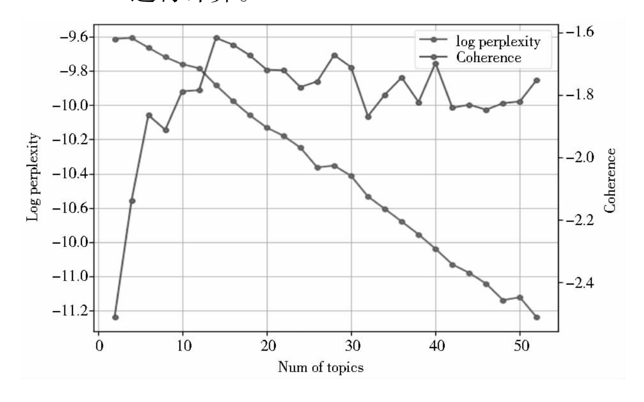


图 4　不同主题个数下的困惑度和连贯性指标

根据图 4 的结果可以看出,随着主题数目的增加,困惑度指标逐渐减小,在 10 几个技术主题的时候,困惑度指标已经趋于稳定,而连贯性指标在 14 - 16 个技术主题时候达到最好,此后趋于稳定甚至下降,根据文献[16],一般来说连贯性指标比困惑度指标评价效果更好,又结合专家对各个主题个数情况下的主题解读情况确定,研究主题的数量选择为 15 个。模型抽取出来的各个主题内容和主题结构见表 2,表中第一列的主题内容标签为向领域专家咨询后给出的,主题结构所在列表示主题中各个关键词及在主题模型中的词语分布权重值。

表 2　干细胞领域研究主题抽取结果

主题序号和内容	主题结构
Topic1 干细胞观测实验	0.036 * "group" + 0.020 * "day" + 0.017 * "study" + 0.016 * "effect" + 0.016 * "control" + 0.016 * "increase" + 0.014 * "result" + 0.014 * "level" + 0.013 * "age" + 0.013 * "compare" + 0.012 * "rat" + 0.011 * "high" + 0.010 * "significantly" + 0.010 * "number" + 0.010 * "method" +0.010 * "week" + 0.009 * "significant" + 0.009 * "change" + 0.009 * "time" + 0.009 * "low"
Topic2 干细胞疾病诊疗	0.019 * "cell" + 0.016 * "disease" + 0.016 * "therapy" + 0.012 * "review" +0.012 * "clinical" + 0.010 * "model" + 0.010 * "stem" + 0.009 * "new" + 0.009 * "approach" + 0.009 * "development" + 0.009 * "therapeutic" +0.008 * "treatment" + 0.008 * "drug" + 0.008 * "base" + 0.008 * "system" +0.007 * "include" + 0.007 * "provide" + 0.007 * "recent" + 0.007 * "study" +0.007 * "discuss"'
Topic3 干细胞癌症治疗	0.063 * "cell" +0.063 * "cancer" +0.052 * "tumor" +0.029 * "msc" +0.014 * "stem" +0.014 * "breast" +0.012 * "expression" +0.011 * "lung" +0.010 * "mesenchymal" +0.010 * "mscs" + 0.010 * "target" + 0.009 * "resistance" + 0.008 * "treatment" + 0.008 * "drug" + 0.008 * "study" + 0.008 * "tumour" + 0.007 * "carcinoma" + 0.007 * "metastasis" + 0.007 * "therapeutic" + 0.007 * "cscs"
Topic4 干细胞与人体组织结构	0.048 * "bone" +0.033 * "tissue" + 0.022 * "scaffold" + 0.015 * "mesenchymal" + 0.014 * "osteogenic" + 0.012 * "regeneration" + 0.012 * "endothelial" + 0.011 * "cell" + 0.011 * "differentiation" + 0.010 * "marrow" +0.010 * "hydrogel" + 0.010 * "vascular" + 0.009 * "cardiac" + 0.009 * "collagen" + 0.009 * "repair" + '0.009 * "study" + 0.008 * "formation" + 0.008 * "cartilage" + 0.008 * "factor" + '0.008 * "derive"
Topic5 干细胞与血液病	0.041 * "patient" + 0.028 * "leukemia" +0.022 * "mutation" + 0.021 * "disease" +0.018 * "aml" + 0.017 * "myeloid" + 0.017 * "case" + 0.014 * "normal" +0.012 * "acute" + 0.010 * "chronic" + 0.009 * "cml" + 0.008 * "leukemic" +0.008 * "disorder" + 0.008 * "associate" + 0.007 * "kit" + 0.007 * "kidney" +0.007 * "blast" + 0.006 * "md" + 0.006 * "syndrome" + 0.006 * "renal"
Topic6 干细胞与神经系统	0.044 * "brain" + 0.041 * "neuron" + 0.029 * "rat" + 0.022 * "neural" +0.020 * "neuronal" + 0.020 * "cell" + 0.015 * "injury" + 0.014 * "adult" +0.013 * "nerve" + 0.013 * "spinal_cord" + 0.012 * "mouse" + 0.007 * "cns" +0.007 * "model" +0.007 * "study" + 0.006 * "motor" + 0.006 * "central_nervous" +0.006 * "stem" + 0.006 * "astrocyte" + 0.006 * "induce" + 0.006 * "follow"
Topic7 干细胞基因作用	0.067 * "gene" + 0.030 * "expression" + 0.023 * "protein" + 0.015 * "sequence" +0.015 * "mrna" + 0.012 * "express" + 0.012 * "rna" + 0.012 * "human" + 0.011 * "virus" + 0.011 * "dna" + 0.010 * "analysis" + 0.010 * "infection" + 0.009 * "level" + 0.009 * "vector" + 0.009 * "clone" + 0.008 * "specific" + 0.008 * "target" + 0.008 * "detect" + 0.007 * "transfer" + 0.007 * "high"
Topic8 干细胞增殖分化	0.185 * "cell" + 0.050 * "stem" + 0.025 * "culture" + 0.023 * "human" +0.020 * "differentiation" + 0.015 * "progenitor" + 0.014 * "derive" +0.010 * "differentiate" + 0.009 * "lineage" + 0.009 * "population" +0.008 * "induce" + 0.007 * "pluripotent" + 0.007 * "factor" + 0.007 * "express" +0.007 * "embryonic" + 0.006 * "potential" + 0.006 * "marker" + 0.006 * "type" +0.006 * "study" + 0.005 * "generate"
Topic9 干细胞作用中的蛋白质功能分析	0.020 * "nucleus" + 0.019 * "neuron" + 0.013 * "bind" + 0.012 * "region" +0.009 * "protein" + 0.009 * "receptor" + 0.008 * "cell" + 0.008 * "site" +0.007 * "contain" + 0.007 * "structure" + 0.006 * "domain" + 0.006 * "response" +0.006 * "terminal" + 0.006 * "label" + 0.006 * "suggest" + 0.006 * "complex" +0.005 * "find" + 0.005 * "activity" + 0.005 * "projection" + 0.005 * "type"
Topic10 干细胞细胞因子层面增殖研究	0.039 * "cell" + 0.022 * "beta" + 0.021 * "expression" + 0.020 * "factor" +0.019 * "induce" + 0.019 * "alpha" + 0.018 * "effect" + 0.015 * "receptor" +0.014 * "growth" + 0.014 * "increase" + 0.013 * "protein" + 0.012 * "pathway" +0.011 * "proliferation" + 0.010 * "level" + 0.010 * "signal" + 0.010 * "role" +0.009 * "activity" + 0.009 * "activation" + 0.009 * "differentiation" +0.009 * "inhibit"
Topic11 细胞表面接触、干细胞培养	0.022 * "cell" + 0.016 * "stem" + 0.012 * "plant" + 0.011 * "activity" + 0.010 * "high" + 0.007 * "root" + 0.007 * "growth" + 0.007 * "show" +0.007 * "leaf" + 0.007 * "concentration" + 0.006 * "wall" + 0.006 * "increase" +0.006 * "membrane" + 0.006 * "surface" + 0.005 * "low" + 0.005 * "property" +0.005 * "study" + 0.005 * "result" + 0.005 * "different" + 0.005 * "response"
Topic12 干细胞与人体组织相关	0.073 * "cell" + 0.019 * "liver" + 0.017 * "tissue" + 0.014 * "epithelial" +0.012 * "culture" + 0.011 * "human" + 0.010 * "stem" + 0.010 * "muscle" +0.009 * "skin" + 0.008 * "expression" + 0.008 * "epithelium" + 0.008 * "study" +0.007 * "regeneration" + 0.007 * "stain" + 0.007 * "hepatocyte" +0.005 * "mouse" +0.007 * "intestinal" + 0.006 * "type" + 0.005 * "marker" + 0.005 * "day"

chinaXiv:202304.00266v1

ChinaXiv合作期刊

（续表 2）

主题序号和内容	主题结构
Topic13 干细胞 基因表达	0.028 * "gene" + 0.017 * "development" + 0.016 * "cell" + 0.016 * "expression" +0.015 * "mouse" +0.011 * "role" +0.010 * "dna" + 0.009 * "mir" + 0.009 * "function" + 0.009 * "factor" + 0.008 * "embryonic" + 0.008 * "embryo" + 0.008 * "mutant" +0.008 * "transcription" + 0.007 * "regulation" + 0.007 * "protein" +0.007 * "regulate" + 0.007 * "early" + 0.006 * "specific" + 0.006 * "mechanism"
Topic14 干细胞 与疾病治疗	0.062 * "patient" + 0.020 * "transplantation" + 0.015 * "stem" + 0.014 * "cell" + '0.013 * "treatment" + 0.013 * "dose" + 0.013 * "chemotherapy" + 0.013 * "high" + 0.011 * "transplant" + 0.011 * "disease" + 0.011 * "survival" + 0.010 * "year" +0.010 * "therapy" + 0.010 * "autologous" + 0.009 * "follow" + 0.009 * "day" + 0.009 * "receive" + 0.009 * "study" + 0.008 * "month" + 0.008 * "relapse"
Topic15 造血干细胞	0.078 * "cell" + 0.060 * "cd" + 0.031 * "marrow" + 0.025 * "blood" + 0.024 * "hematopoietic" +0.023 * "bone" +0.016 * "mouse" +0.016 * "progenitor" + 0.011 * "peripheral" + 0.011 * "csf" + 0.009 * "number" + 0.008 * "stem" + 0.007 * "day" + 0.007 * "donor" + 0.006 * "transplantation" + 0.006 * "factor" + 0.006 * "antigen" + 0.006 * "cytokine" + 0.006 * "platelet" + 0.006 * "gm_csf"

3.3　主题强度计算及关联性分析

（1）3 种相似度度量方法结果。根据公式（3）（4）（5）分别计算，得到不同相似性度量下的各个研究主题的相似性矩阵，以热力图的形式表示各个研究主题之间的相似性（见图 5），图中颜色越深的部分表示相似度数值越大。

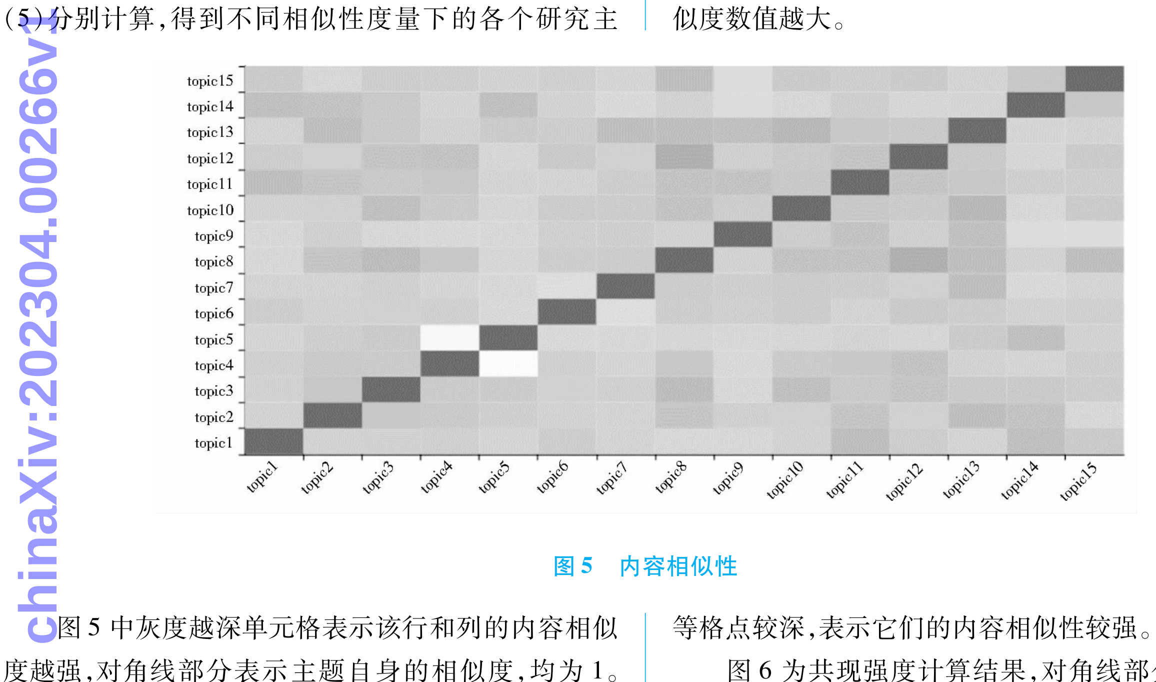


图 5　内容相似性

图 5 中灰度越深单元格表示该行和列的内容相似度越强，对角线部分表示主题自身的相似度，均为 1。内容相似度具有对称性，图中表现为关于对角线对称的格点深浅相同。图中 topic8-topic12、topic10-topic13 等格点较深，表示它们的内容相似性较强。

图 6 为共现强度计算结果，对角线部分为主题自身共现强度，即主题出现的频次。共现强度也具有对称性。从图 6 可以看出 topic8-topic13、topic8-topic10 等

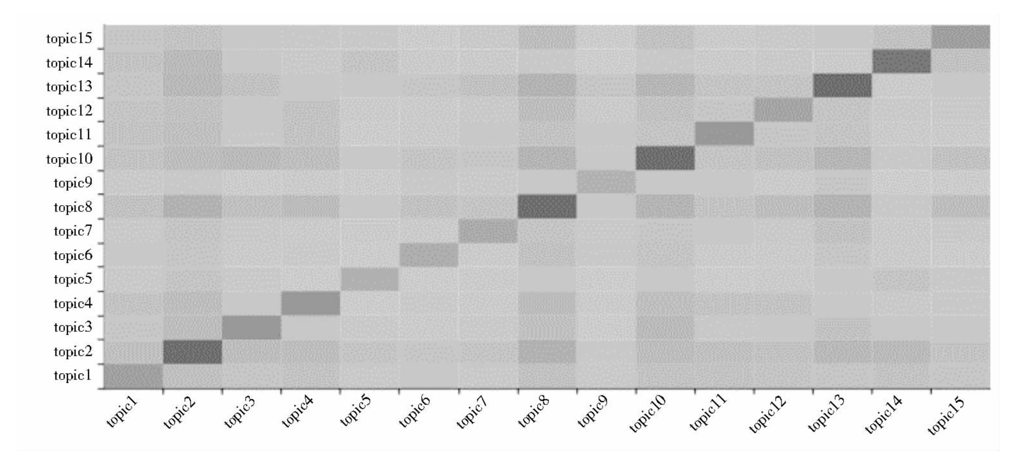


图 6　共现强度

chinaXiv:202304.00266v1

ChinaXiv合作期刊

对应单元格较深，说明这些主题共同出现频次高，本身出现频次较高的主题（如 topic8、topic10 等）与其他主题共现强度也相对高些。

图 7 为趋势相似性计算结果，对角线部分为主题自身的趋势相似性为 1。趋势相似性也具有对称性，关于对角线对称的格点颜色深浅相同。从图 7 可以看出 topic5-topic6、topic3-topic4 格点较深，说明这些主题具有相似的趋势。

（2）3 种相似性度量方法的一致性分析结果。

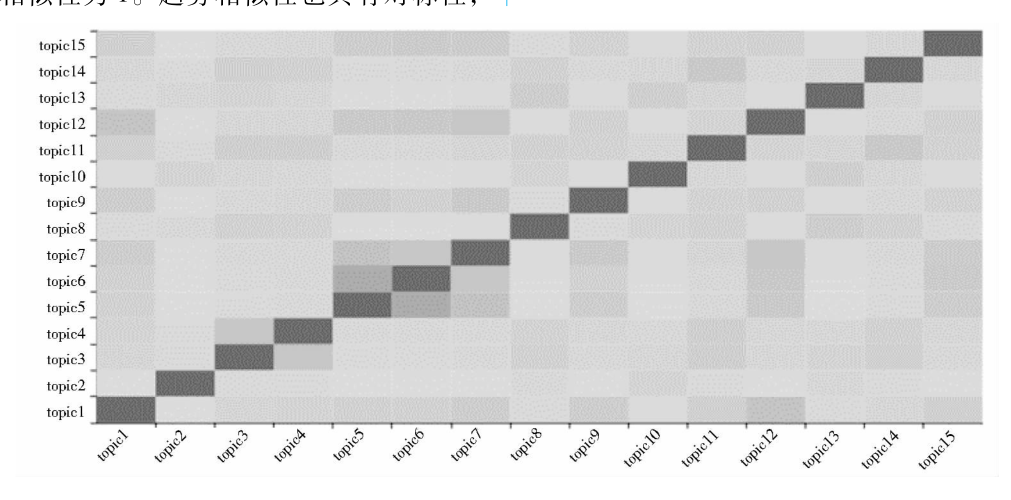


图 7　趋势相似性

表 3　一致性分析结果

主题	内容相似性	共现相似性	趋势相似性	内容－趋势	共现－趋势	内容－共现
Topic1	[1, 11, 14, 15, 12, 6, 4, 3, 10, 2, 7, 13, 5, 8, 9],	[1, 2, 8, 4, 11, 10, 14, 12, 13, 15, 3, 6, 7, 9, 5],	[1, 12, 7, 9, 11, 15, 5, 6, 4, 3, 14, 8, 13, 10, 2],	11	13	12
Topic2	[2, 13, 14, 11, 8, 3, 4, 5, 9, 10, 12, 6, 1, 7, 15],	[2, 8, 13, 14, 10, 4, 3, 11, 1, 12, 15, 5, 7, 6, 9],	[2, 10, 13, 8, 3, 4, 11, 14, 1, 9, 15, 12, 7, 6, 5],	11	8	11
Topic3	[3, 8, 10, 12, 2, 14, 11, 13, 4, 5, 15, 7, 1, 6, 9],	[3, 10, 2, 8, 13, 1, 4, 14, 15, 12, 11, 7, 5, 6, 9],	[3, 4, 11, 14, 8, 13, 1, 10, 12, 15, 9, 7, 6, 2, 5],	12	10	9
Topic4	[4, 12, 11, 8, 2, 10, 3, 15, 6, 1, 13, 7, 14, 9, 5],	[4, 8, 2, 10, 1, 11, 12, 3, 13, 15, 14, 6, 7, 9, 5],	[4, 3, 11, 14, 8, 1, 13, 12, 15, 9, 10, 7, 6, 5, 2],	11	11	9
Topic5	[5, 14, 13, 3, 2, 15, 11, 1, 10, 12, 8, 9, 7, 6, 4],	[5, 2, 14, 10, 13, 8, 15, 3, 1, 12, 7, 4, 11, 6, 9],	[5, 6, 7, 12, 15, 9, 1, 11, 4, 3, 14, 13, 8, 10, 2],	13	14	10
Topic6	[6, 12, 10, 8, 13, 1, 9, 4, 15, 3, 2, 14, 11, 5, 7],	[6, 8, 10, 2, 13, 1, 9, 4, 11, 12, 3, 14, 7, 15, 5],	[6, 5, 7, 12, 15, 9, 1, 11, 4, 3, 14, 8, 13, 10, 2],	12	11	8
Topic7	[7, 13, 10, 8, 11, 9, 3, 12, 1, 2, 4, 15, 5, 14, 6],	[7, 13, 2, 8, 10, 11, 15, 1, 3, 9, 12, 4, 14, 5, 6],	[7, 5, 6, 12, 9, 15, 1, 11, 4, 3, 14, 13, 8, 10, 2],	13	13	9
Topic8	[8, 12, 3, 13, 15, 11, 10, 2, 4, 7, 6, 9, 14, 5, 1],	[8, 2, 13, 10, 4, 15, 12, 3, 1, 6, 11, 7, 5, 9, 14],	[8, 13, 3, 14, 11, 4, 10, 2, 1, 15, 9, 12, 7, 6, 5],	11	13	12
Topic9	[9, 13, 11, 10, 7, 6, 2, 12, 8, 5, 3, 1, 4, 15, 14],	[9, 13, 11, 10, 8, 2, 6, 1, 7, 12, 4, 3, 15, 5, 14],	[9, 7, 5, 1, 6, 12, 15, 11, 4, 3, 14, 8, 13, 10, 2],	13	13	9
Topic10	[10, 13, 3, 8, 11, 15, 4, 12, 7, 6, 9, 2, 1, 5, 14],	[10, 8, 13, 3, 2, 4, 15, 12, 1, 11, 6, 7, 5, 9, 14],	[10, 13, 2, 8, 3, 11, 14, 4, 1, 15, 9, 12, 7, 6, 5],	10	12	10
Topic11	[11, 1, 9, 8, 2, 12, 13, 4, 10, 3, 7, 14, 15, 6, 5],	[11, 2, 8, 1, 4, 10, 13, 12, 9, 7, 3, 6, 15, 14, 5],	[11, 14, 3, 4, 1, 8, 9, 15, 12, 13, 7, 10, 6, 5, 2],	12	11	11
Topic12	[12, 8, 4, 11, 3, 13, 6, 10, 15, 1, 9, 7, 2, 5, 14],	[12, 8, 10, 2, 4, 13, 1, 11, 3, 15, 7, 14, 9, 6, 5],	[12, 1, 7, 6, 5, 15, 9, 11, 4, 3, 14, 8, 13, 10, 2],	14	13	11
Topic13	[13, 10, 8, 7, 2, 9, 11, 12, 3, 5, 6, 4, 15, 1, 14],	[13, 8, 10, 2, 3, 7, 12, 11, 9, 1, 6, 15, 5, 4, 14],	[13, 8, 10, 3, 11, 14, 4, 2, 1, 15, 9, 12, 7, 6, 5],	14	11	10
Topic14	[14, 1, 5, 2, 15, 3, 11, 8, 6, 4, 13, 12, 10, 7, 9],	[14, 2, 15, 1, 5, 3, 10, 8, 4, 12, 7, 13, 11, 6, 9],	[14, 11, 3, 4, 8, 13, 1, 15, 10, 9, 12, 7, 6, 5, 2],	14	13	11
Topic15	[15, 8, 14, 10, 12, 1, 3, 11, 4, 6, 5, 13, 7, 2, 9],	[15, 8, 10, 14, 2, 1, 13, 3, 5, 4, 7, 12, 11, 6, 9],	[15, 6, 7, 5, 1, 12, 9, 11, 4, 14, 3, 8, 13, 10, 2],	11	13	10
平均值				12.13	11.93	10.13

chinaXiv:202304.00266v1

ChinaXiv合作期刊

表 3 的最后一列是对 15 个研究主题的编辑距离计算的平均值。从表 3 可以明显看出：①几种相似性判别公式之间的编辑距离都比较大，说明内容相似性、共现相似性、趋势相似性 3 种关联判别方法之间的一致性不强，内容/共现相似性较强的研究主题不一定具有一致的发展趋势；②3 种相似性判别方法的一致性顺序关系为：(内容，共现) > (共现，趋势) > (内容，趋势)。

3.4　主题强度演化分析与预测

笔者采用 3 个指标衡量观测值与真实值之间的偏差，评价趋势预测结果的好坏。

RMSE 均方根误差的计算公式如下所示：

$$RMSE(ST_i, \hat{ST}_i) = \sqrt{\frac{1}{m}\sum_{u=1}^{U}(ST_{ui} - \hat{ST}_{ui})^2}$$

公式（12）

MAE 平均绝对值误差的计算公式如下所示：

$$MAE(ST_i, \hat{ST}_i) = \frac{1}{m}\sum_{u=1}^{U}|ST_{ui} - \hat{ST}_{ui}|$$

公式（13）

R2Score 决定系数(拟合优度)，R2 值取值范围是 0 到 1，R2 越接近 1，说明拟合效果越好。计算公式如下所示：

$$R2(ST_i, \hat{ST}_i) = 1 - \frac{\sum_{i=1}^{U}(ST_{ui} - \hat{ST}_{ui})^2}{\sum_{i=1}^{U}(ST_{ui} - \bar{ST}_{ui})^2}$$

公式（14）

表 4 中给出了 3 种趋势预测方法的结果比较，第 1 列表示采用 ARMIA 时间序列分析模型(用 Auto－ARIMA 工具)分析计算得到的结果，第 2 列为采用 LSTM 时间序列分析模型得到的结果，后面 3 列表示用 Prophet 模型进行序列预测的结果，其中，第 3 列表示采用通常的先取对数，拟合模型后再还原的方式，第 4 列表示直接对原始数据采用线性趋势进行预测的结果，第 5 列表示直接对原始数据采用 Logistic 趋势进行预测的结果，其中每个单元格中数字为(RMSE，MAE，

表 4　几种趋势预测方法的结果对比

主题	Auto－ARIMA	LSTM 方法	Prophet－原始数据取对数再还原	Prophet－线性趋势预测	Prophet-Logistic 趋势预测
Topic1	(88.807 7,75.336 9,0.886 5)	(195.722 1, 116.318,0.896)	(102.258 3,72.564 1,0.975 1)	(41.700 1,29.022 0,0.995 8)	**(35.434 7,26.915 8,0.997 0)**
Topic2	(459.14 35,287.087 3,0.466 1)	(398.211, 235.832 1,0.9182)	(284.142 3,185.118 1,0.960 6)	(461.674 5,408.370 6,0.896 0)	**(167.939 1,102.151 3,0.986 2)**
Topic3	(341.466 8,291.258 5,0.409 3)	(386.798 8, 215.683 4,0.823 4)	(167.013 8,112.866 4,0.968 0)	(361.574 3,318.511 4,0.850 0)	**(116.937 8,66.851 7,0.984 3)**
Topic4	(391.983 2,199.609 9,0.025 6)	(314.041 1, 188.623 2,0.870 7)	(151.028 3,93.007 9,0.971 1)	(300.529 8,267.735 3,0.885 6)	**(109.884 3,58.640 7,0.984 7)**
Topic5	(39.440 0,32.421 1,0.949 6)	(66.090 9, 46.815 8,0.969 7)	(86.640 6,62.010 8,0.958 1)	**(20.419 8,16.830 2,0.997 6)**	(31.952 7,24.230 2,0.994 3)
Topic6	(329.252 1,297.254 2,－3.780 2)	(113.606 2, 74.666 3,0.924 8)	(83.143 6,56.430 9,0.966 8)	(28.160 1,22.368 9,0.996 2)	**(26.512 5,20.522 8,0.996 6)**
Topic7	(77.227 2,63.228 0,0.843 7)	(102.012 2, 64.831 5,0.923 4)	(109.985 8,73.675 6,0.9350 4)	**(48.081 1,33.767 6,0.987 6)**	(51.719 5,32.332 5,0.985 6)
Topic8	(1 060.000 7,824.576 6,－24.162 3)	(367.57, 235.049,0.871 4)	(186.348 6,140.179 5,0.970 3)	(263.459 4,235.300 7,0.940 7)	**(82.334 3,64.098 5,0.994 2)**
Topic9	(347.405 1,224.650 4,－0.226 6)	(88.474 7, 55.711 7,0.902 4)	(167.470 5,82.755 4,0.816 2)	(150.376 2,87.840 3,0.851 8)	**(121.841 3,62.114 8,0.902 7)**
Topic10	(330.003 3,230.773 4,0.640 9)	(367.477 8, 220.741,0.902)	(206.935 8,142.639 8,0.970 7)	(407.734,361.183 3,0.886 6)	**(120.958 6,82.393 7,0.990 0)**
Topic11	(93.397 6,82.227 2,0.942 7)	(260.592 7, 147.442 2,0.865 4)	(145.996 0,106.102 2,0.964 8)	**(38.936 4,30.761 8,0.997 5)**	(50.119 2,35.645 5,0.995 8)
Topic12	(153.775 2,104.745 1,0.425 7)	(137.881 2, 88.926 8,0.930 5)	(96.586 1,67.529 7,0.971 1)	**(23.308 4,18.152 7,0.998 3)**	(27.502 3,20.341 0,0.997 7)
Topic13	(912.392 7,644.149 7,－5.287 7)	(411.129 2, 241.578 5,0.862 9)	(201.983 1,143.382 6,0.969 1)	(343.228 2,307.743 1,0.910 7)	**(118.928 5,77.828 9,0.989 3)**
Topic14	(243.774 6,217.726 5,0.160 1)	(134.401 1, 100.039 1,0.968 8)	(248.225 8,187.647 5,0.911 5)	(117.263 7,85.140 9,0.980 2)	**(96.975 6,80.465 8,0.986 5)**
Topic15	(193.582 4,116.529 1,0.272 3)	(113.243 9, 82.245 2,0.910 9)	(152.298 3,99.399 7,0.894 5)	(94.092 2,62.707 9,0.959 7)	**(89.104 9,56.290 9,0.963 9)**

chinaXiv:202304.00266v1

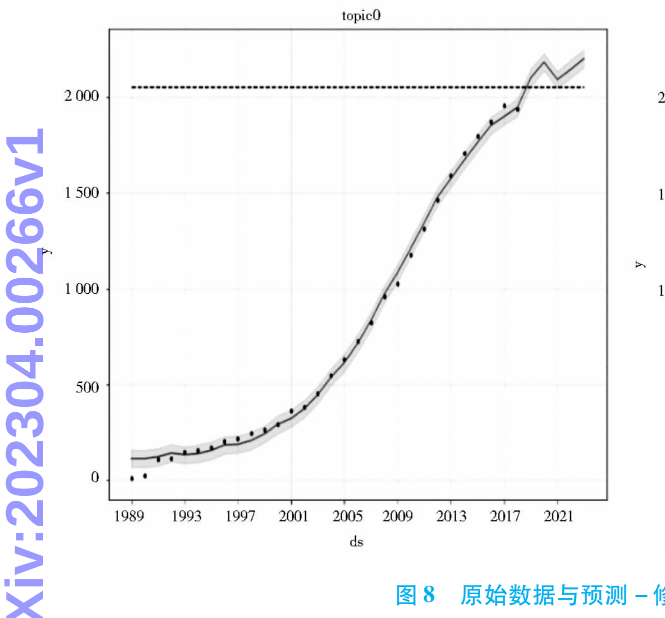
ChinaXiv合作期刊

R2Score)值. 从上表可以看出如下结论：

(1)对各个研究主题而言,Prophet 模型的 R2Score 值都达到0.90 以上,比 ARIMA 模型和 LSTM 模型效果要好些,说明 Prophet 模型能够很好地拟合研究主题的演化趋势。对原始数据取对数再还原的方式并没有提高预测的准确性。由于各个主题的分布具有明显的增长趋势,序列是非平稳的,直接使用 ARIMA 模型效果较差。又由于本例中以年为时间片,数据相对较少,LSTM 模型训练不充分,效果不是很理想,而且易发生过拟合现象。而 Prophet 模型中的 Logistic 增长模式与该领域研究主题的增长模式相符合,且模型参数较少,容易达到较好的效果。

(2)各个研究主题的时间演化规律并不一致,大部分的研究主题更加符合 Logistic 趋势,但 Topic5、Topic7、Topic11 与 Topic12 常采用线性趋势拟合效果更佳,略好于 Logistic 模型,可能由于这几个研究主题正处在高速增长时期,尚未达到饱和增长。

原始预测模型与预测 – 修正模型结果对比如图 8 所示：



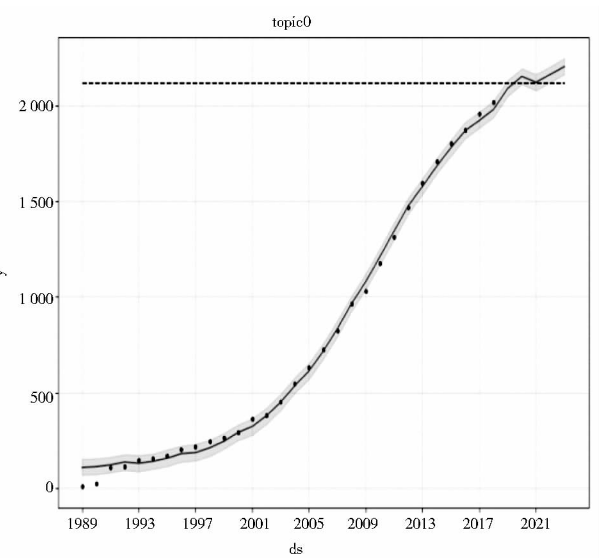


图 8　原始数据与预测 – 修正数据的拟合结果

图 8 为同一个技术主题的预测结果,左图是对 2018 年的原始数据进行拟合与预测结果,右图是经过预测 – 修正之后的拟合预测结果,显然在预测 – 修正模型使得后面趋势预测部分相对更平稳,波动性更小,也更符合主题演化增长规律。

3.5　主题强度演化分析与预测结果展示

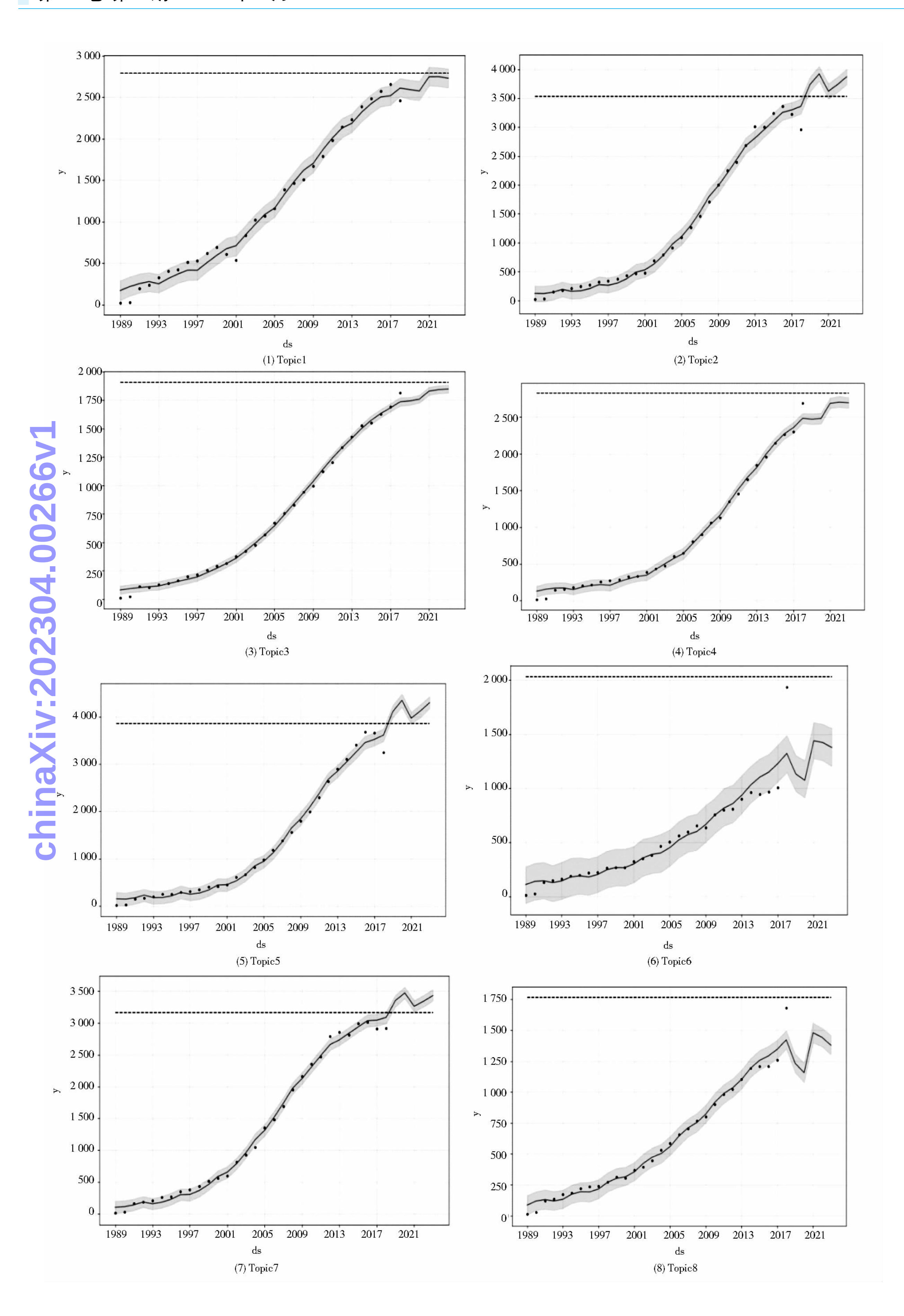
使用基于 Prophet 的预测 – 修正模型,在 15 个研究主题上分别进行模型拟合,得到结果见图 9。从图中可以看出,各个研究主题基本符合模型增长规律,但置信区间宽度区别比较大。异常值数据,特别是最近时间片的异常数据对后面预测结果影响比较大,可以使得后续模型预测波动变大,如 topic2、topic10 等;此外异常数据还可能使得置信区间变得更宽,如 topic6、topic15 等。

采用主题河流图的方式将各个技术主题的强度演化趋势进行展示,在图 10 中将各个研究主题的趋势分析图进行了叠加,每一个灰度条代表一个研究主题,灰度条的宽度代表不同研究主题的主题强度,图中可以明显看出不同研究主题强度对比于强度演化关系。图 10(1)为未引入预测模型后对原始数据的展示,图 10(2)和 10(3)为引入基于 Prophet 的预测 – 修正模型之后的结果,其中 10(2)中灰度条的宽度代表主题强度,图 10(3)中灰度条的宽度代表主题相对强度,即该时间片内特定主题的强度与全部主题强度之和的比值,主题相对强度更能反应领域内各个主题的结构性变化。右侧方框部分为 2018 年之后的结果,图 10(2)和 10(3)的效果相对更符合演化的发展规律,很大程度上解决了近期数据不完整对趋势分析的影响。

从图 10(2)中可以看出,干细胞各个研究主题整体呈现增长态势。但从图 10(3)可以看出领域中的各个研究主题变化趋势还是有区别的,如 Topic2(干细胞与疾病诊疗)这个研究主题发展态势最为迅猛,在整体研究中所占的比例逐渐增大,干细胞在越来越多的疾病诊疗中得到应用。而如 Topic15(造血干细胞)等研究主题,理论研究已经相对较为成熟,在整体研究中占的比例呈现下降趋势,某些成熟技术已经从理论走向应用。

chinaXiv:202304.00266v1

ChinaXiv合作期刊



(1) Topic1 (2) Topic2 (3) Topic3 (4) Topic4 (5) Topic5 (6) Topic6 (7) Topic7 (8) Topic8

chinaXiv:202304.00266v1

chinaXiv:202304.00266v1

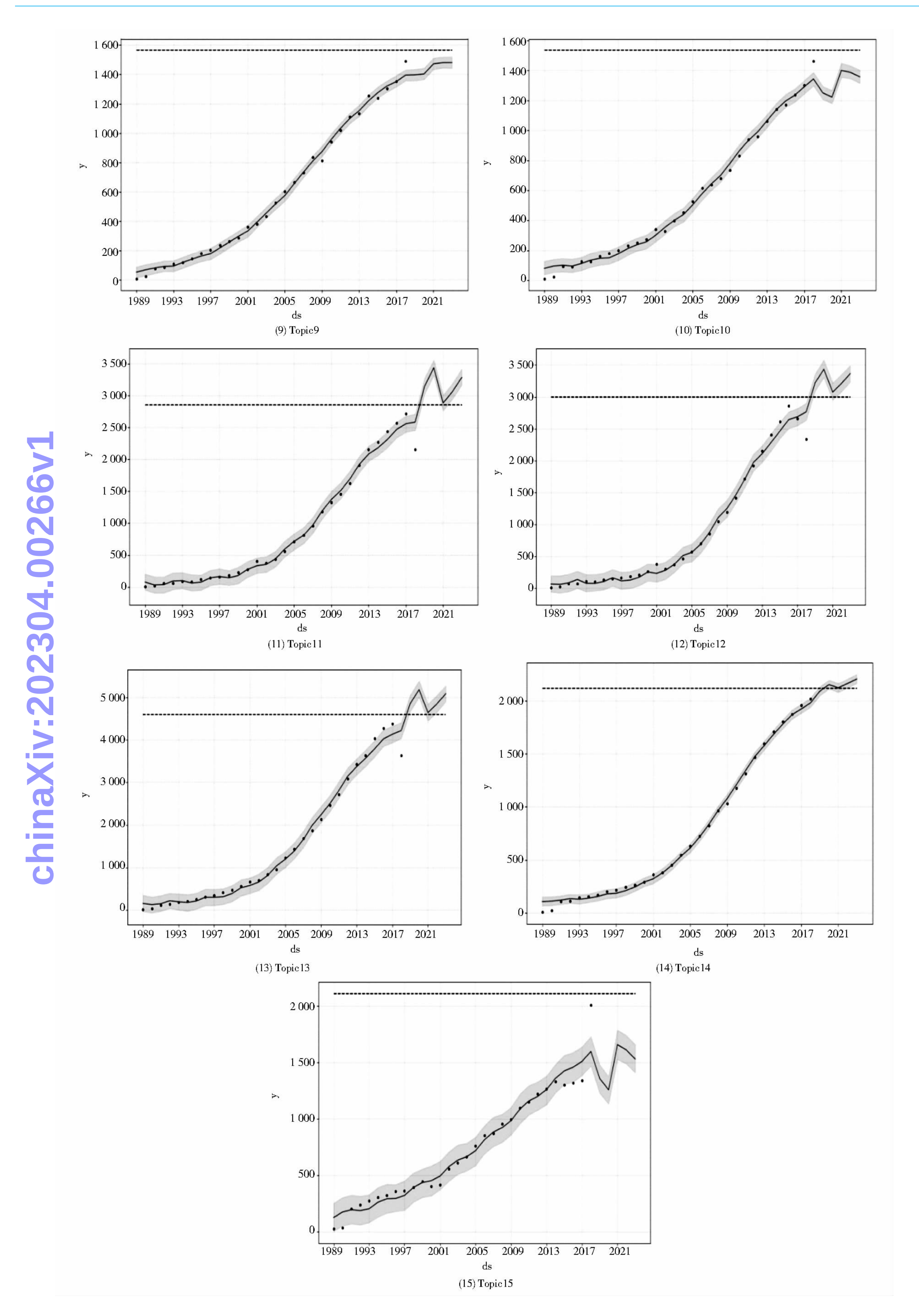


图 9　每个研究主题的演化趋势

ChinaXiv合作期刊

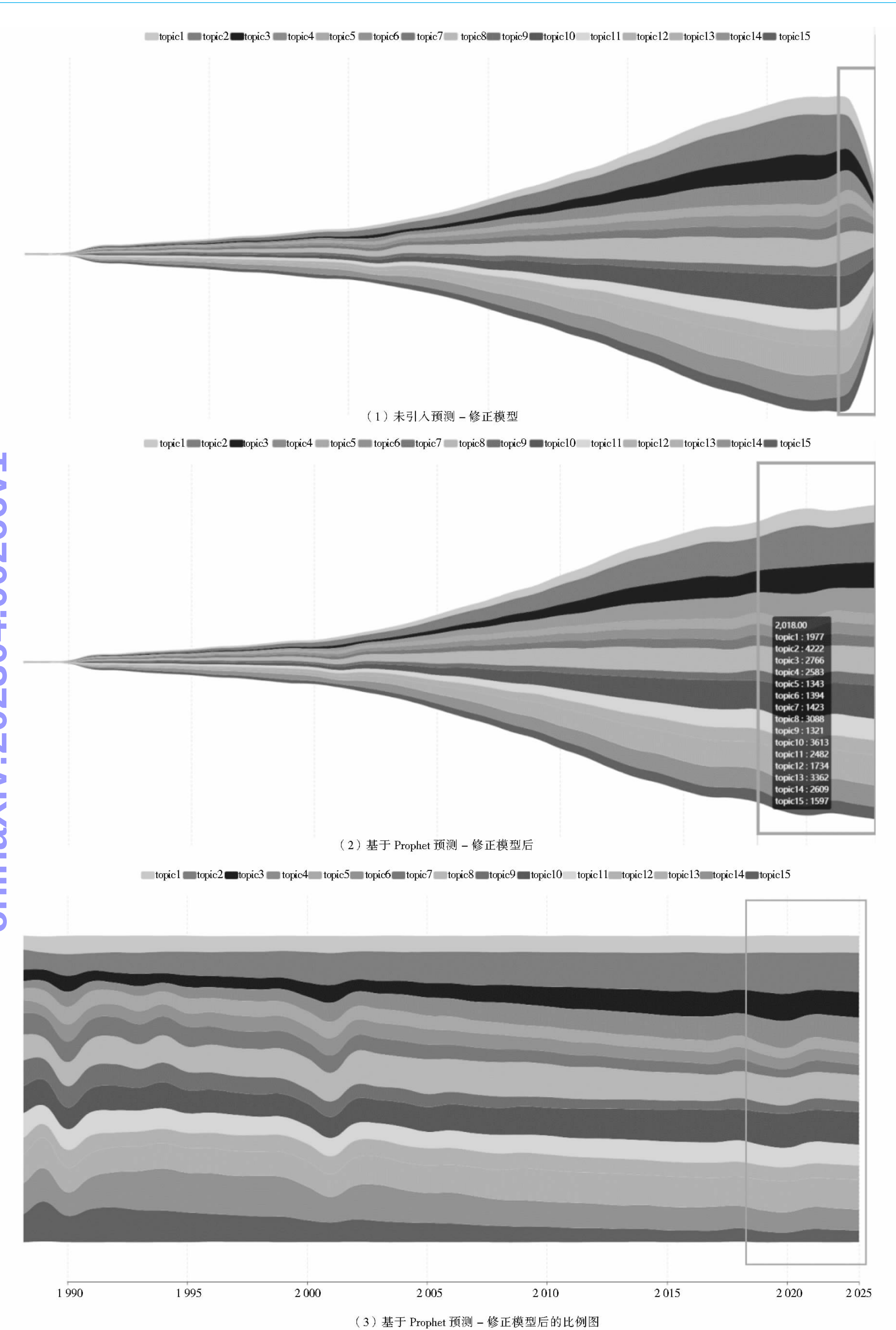


图 10　主题强度趋势演化主题河流图展示

chinaXiv:202304.00266v1

ChinaXiv合作期刊

在1990年与2000年前后,从图10(2)中可以看出各个研究主题增长趋势呈现明显变化(类似曲线中的拐点),从图10(3)中看出各个研究主题分布也呈现震荡趋势。这两个时间点与1988年詹姆斯汤姆森分离出人类胚胎干细胞等重大发现,以及1999年美国将干细胞列为十大科学进展,2000年日本干细胞计划提出大致相符,说明重大科学发现不仅可能会带来主题增长趋势变化,也可能会带来主题结构的变化。

4 总结讨论

研究主题分析是进行科技决策的基础,目前对研究主题的趋势分析大多专注于研究主题的态势描述,对研究主题趋势规律性分析和预测研究的不够充分。笔者提出一种研究主题的趋势分析和预测方法,整体模型分为主题抽取和表示、主题关联和相似性分析、主题趋势分析和预测。在主题抽取和表示阶段,主要使用LDA模型进行抽取表示。在主题关联和相似性分析阶段,使用通常的内容相似性、以及笔者提出的共现相似性、趋势相似性等指标进行主题相似性度量,并探讨几种主题相似性度量方法的相互关系,得到几种方式的一致性关系为(内容,共现)>(共现,趋势)>(内容趋势)。在主题的趋势分析和预测阶段,将Prophet模型引入到主题演化分析当中,进行强度演化趋势分析与预测,比较了Prophet模型与ARIMA、LSTM等经典模型的效果。针对强度演化中的近期数据不完整问题,笔者提出了的预测－修正的两阶段模型,实验证明该模型能够较好地拟合主题现状和未来演化趋势。

模型的优缺点与后续工作如下:

(1)在干细胞领域进行实证分析,数据已经达到了几十万的规模,时间分片后主题呈现明显的趋势特征,可以采用时间序列模型进行拟合。但数据覆盖范围仍不够全面,未来将考虑在选择其它的典型领域进行实证,进行不同领域数据的对比分析,进一步验证笔者提出方法的适用性。

(2)在主题抽取与表示阶段,使用LDA模型进行研究主题的表示,LDA模型具有能够自动抽取文档隐含主题表示等诸多优点,也是目前主题抽取表示中的经典方法,但LDA主题模型存在表示能力与可解释性不够强等问题,还需要领域专家对研究主题进行相应的解读,且LDA模型假设不同时间片内的主题数量是一致的,这也是目前主题演化实用系统常用的处理方式,这样处理虽然有计算量小等优势,但在表示新兴主题产生面有些不足,后续可以考虑结合预训练模型等方法进行研究主题表示模型的改进。

(3)在主题强度趋势预测中,使用Prophet预测－修正模型进行拟合分析,具有自动化程度高、易于实现,与主题增长规律符合等优点。本文将年作为时间切片的单位,以1年为一个时间单元,这样处理比通常情报学文章选取的3－5年为时间片细一些,更能体现趋势特征。但仍旧力度较粗,未充分利用Prophet的周期和节假日模型的丰富表示能力,未来可以使用更大的数据、在更细的粒度进行趋势预测和分析。

(4)通过不同研究主题符合的不同增长模式,对研究主题进行归类,进行新兴与热门研究主题挖掘也是后续值得研究的问题。可以通过主题演化趋势曲线中的拐点或震荡点,来推断重大发现或颠覆性技术出现的时间。

参考文献:

[1] 罗文馨,王园园. 技术主题演化研究方法综述[J]. 知识管理论坛,2018,3(5):255－265.
[2] HUMMON N P, DEREIAN P. Connectivity in a citation network: the development of DNA theory[J]. Social networks, 1989, 11(1): 39－63.
[3] MARTINELLI A. An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry[J]. Research policy, 2012, 41(2): 414－429.
[4] LU L Y Y, LIU J S. A survey of intellectual property rights literature from 1971 to 2012: the main path analysis[C]//Proceedings of PICMET'14 conference: portland international center for management of engineering and technology; infrastructure and service integration. Piscataway: IEEE, 2014: 1274－1280.
[5] PILKINGTON A, MEREDITH J. The evolution of the intellectual structure of operations management－1980－2006: a citation/co-citation analysis[J]. Journal of operations management, 2009, 27(3): 185－202.
[6] LAI R J, LI M F. Technology evolution of lower extremity exoskeleton from the patent perspective[J]. Key engineering materials. 2014, 625: 536－541.
[7] WANG Z Y, LI G, LI C Y, et al. Research on the semantic-based co-word analysis[J]. Scientometrics,2012,90(3):855－875.
[8] 胡正银,刘春江,隗玲,等. 面向TRIZ的领域专利技术挖掘系统设计与实践[J]. 图书情报工作,2017,61(1):117－124.
[9] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993－1022.
[10] 范少萍,安新颖,单连慧,等. 基于医学文献的主题演化类型与演化路径识别方法研究[J]. 情报理论与实践,2019,42(3):114－119.
[11] BLEI D M, LAFFERTY J D. Dynamic topic models[C]//Pro-

chinaXiv:202304.00206v1

ChinaXiv合作期刊

ceedings of the 23rd international conference on Machine learning. New York:ACM, 2006: 113 – 120.
[12] WANG X, MCCALLUM A. Topics over time: a non-Markov continuous-time model of topical trends[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York:ACM, 2006: 424 – 433.
[13] PORTEOUS I, NEWMAN D, IHLER A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York:ACM, 2008: 569 – 577.
[14] HOFFMAN M, BACH F R, BLEI D M. Online learning for latent dirichlet allocation[C]//Advances in neural information processing systems 23. Vancouver:Curran Associates Inc., 2010: 856 – 864.
[15] GRIFFITHS T L, JORDAN M I, TENENBAUM J B, et al. Hierarchical topic models and the nested Chinese restaurant process [C]//Advances in neural information processing systems. Vancouver:ACM,2004: 17 – 24.
[16] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing semantic coherence in topic models[C]//Proceedings of the conference on empirical methods in natural language processing. Edinburgh: Association for Computational Linguistics, 2011: 262 – 272.
[17] 王婷婷,韩满,王宇. LDA 模型的优化及其主题数量选择研究——以科技文献为例[J]. 数据分析与知识发现,2018,2(1):29 – 40.
[18] WANG X, MCCALLUMA A, WEI X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval [C]// IEEE International Conference on Data Mining. Piscataway: IEEE,2007:697 – 702.
[19] LI B, WANG B, ZHOU R, et al. CITPM: A cluster-based iterative topical phrase mining framework[C]//International conference on database systems for advanced applications. Dallas: Springer International Publishing,2016:197 – 213.
[20] 张琴,张智雄. 基于 PhraseLDA 模型的主题短语挖掘方法研究[J]. 图书情报工作,2017,61(8):120 – 125.
[21] 刘自强,许海云,岳丽欣,等. 基于 Chunk-LDAvis 的核心技术主题识别方法研究[J]. 图书情报工作,2019,63(9):73 – 84.
[22] 孙孟孟. 基于名词短语提取与词条权重分析的话题提取算法研究[D]. 杭州:浙江大学,2014.
[23] GRAVES A. Supervised sequence labelling [M]//Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012: 5 – 13.
[24] TAYLOR S J, LETHAM B. Forecasting at scale[J]. The American statistician, 2018, 72(1):37 – 45.

作者贡献说明:
张鑫:提出研究思路,撰写论文;
文奕:提出研究问题,修改论文;
许海云:修改论文;
刘忠禹:模型结果评价。

chinaXiv:202304.00266v1

Prophet Prediction-Correction Topic Evolution Model——A Case Study in Stem Cell Field

Zhang Xin　Wen Yi　Xu Haiyun　Liu Zhongyu
Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041

Abstract: [**Purpose/significance**] Topic evolution analysis plays an important role in detection the technology frontier detection and innovation strategy deployment. [**Method/process**] In this paper, the topic evolution analysis process was divided into several steps: topic representation, similarity correlation and intensity evolution calculation. The LDA model was used to represent the topic; content, co-occurrence, and trend similarity were proposed for topic correlation calculations, and the prophet-based pre-train fine-tuning model was used to predict the topic trends. An empirical analysis was conducted using the stem cell field as an example. [**Result/conclusion**] Experiments show that the Logistic growth model has a R2Score of more than 0.90 for each topic. It shows that the Logistic growth model in Prophet is consistent with the growth trend of topics, and can fit the evolution trend of the topic intensity. The topic evolution model proposed in this paper has certain reference to topic distribution and evolution analysis in specific fields.

Keywords: topic evolution　topic similarity　time series　Prophet